Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Estudio de los métodos de análisis multivariante para la detección de la tendencia de compra en clientes

MEMORIA

Autor: Sergio González Marinas
Director: Josep Anton Sànchez Espigares
Convocatoria: abril 2020

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resumen

El siguiente proyecto trata sobre el estudio y aplicación práctica de dos de los tipos más utilizados de análisis multivariante. La aplicación práctica es sobre una matriz de datos proporcionada por un proveedor de artículos de peluquería y cosméticos. Para la aplicación práctica se utilizará el *software* R™.

El proyecto comienza con una introducción teórica a los modelos y métodos matemáticos necesarios para entender el funcionamiento de los dos métodos de análisis multivariante: el análisis de componentes principales y el análisis factorial. Posteriormente, se expondrán y analizarán los resultados obtenidos de la aplicación de estos métodos con el programa R™ y se discutirá si es interesante el uso de estos análisis para obtener una tendencia en la compra de los clientes, y así, poder implementar métodos de actuación para la retención de clientes y/o mejoras en la facturación de la empresa. En conclusión, el trabajo busca discutir si el uso de los métodos de análisis multivariante son óptimos para observar la tendencia de los clientes, concretamente los clientes de un proveedor de productos de peluquería.

Sumario

1. Glosario

ACP: Análisis de componentes principales.

AF: Análisis factorial

AFE: Análisis factorial exploratorio.

AM: Análisis multivariante.

Análisis clusters: Análisis de conglomerados

Barplot: Diagrama de barras.

2. Introducción

En el paradigma del mercado actual, poder saber la tendencia de compra de los clientes es una inversión muy buscada e interesante para los modelos de empresas contemporáneas. Poder decidir los planes de acción con el respaldo de datos y modelos es algo muy necesario en algunas empresas que manejan volúmenes de facturación elevados, e incluso, en pequeño y mediano comercio. La implementación de métodos rápidos y fiables pueden ser un punto a favor frente a los competidores del sector [1]. La recogida de datos sobre cada cliente es algo muy recurrente en la mentalidad empresarial actual y con esta recogida masiva surge la problemática del exceso de variables y de abultar tanto la dimensión como el tiempo de análisis. Por lo tanto, los métodos de análisis multivariante son interesantes tanto para reducir la dimensión del problema, reduciendo las variables de entrada, como para poder obtener variables de salida que expliquen esa tendencia de compra que buscan las empresas.

2.1. Objetivos generales del proyecto

Los objetivos generales del proyecto son el propósito que se pretende conseguir mediante la realización del proyecto. En este caso, se pretende que los objetivos cumplan con el criterio SMART, siendo así objetivos específicos, medibles, asequibles, realistas y de duración determinada. Los objetivos son los siguientes:

- Estudiar los métodos de análisis multivariante que más encajan con los datos de entrada, observar y analizar las diferencias entre los dos métodos (ACP y AFE), y discutir cual es más conveniente.

- Analizar si es óptimo utilizar las variables de salida de los métodos de análisis multivariante para caracterizar la tendencia de compra de ciertos clientes.

2.2. Alcance del proyecto

El proyecto que se define tiene unos límites de actuación; temas, aspectos o decisiones donde el proyecto no entrará y que se detallará que quedan excluidos del estudio del proyecto. Los aspectos que quedan excluidos del proyecto son:

- **Cualquier otro método de análisis multivariante fuera de los elegidos.** Ya que se han elegido los dos métodos más convenientes para los datos de entrada, no se cree preciso entrar a discutir otros métodos y variantes ya que daría lugar a un trabajo más extenso y fuera del alcance de este proyecto.

- **Utilizar cualquier otro *software* que no sea R™.** Hay muchos programas disponibles para realizar este proyecto, no obstante, se utilizará el citado *software* ya que es uno de los más utilizados en docencia y estadística.

3. Programación del proyecto

Para una correcta realización de cualquier proyecto a largo plazo se debe estar organizado antes de empezarlo. Esto ayuda a imponer plazos para su entrega a tiempo y para hacerlo lo más óptimo posible se decide realizar un diagrama de Gantt.

3.1. Diagrama de Gantt

Se exponen las siguientes concreciones necesarias para la realización del diagrama de Gantt:

- La unidad temporal serán semanas.
- Un cuadrimestre tiene 18 semanas y el trabajo de final de grado equivale a 12 ECTS, lo cual se traslada a una carga de trabajo de 25 h /ECTS. Por lo tanto, para llegar a esa carga de trabajo se aplica una media semanal de trabajo de 15 h/ semana.
- La estructuración del proyecto es la expuesta en el capítulo del Índice más todo el tiempo dedicado a programar con R y obtener los resultados.

Como se verá en la figura 5.3.1. se puede resumir la programación temporal y la estructuración del proyecto en la siguiente tabla. Esta información es valiosa para el cálculo del presupuesto total del capítulo 6.

Estructura Proyecto Final de Grado	
Semanas invertidas	18 semanas
Horas invertidas	270 horas
Fecha inicio / finalización	16 de diciembre / 14 de abril

Actividad	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
Fundamentos teóricos																		
Programación en R																		
Análisis resultados																		
Presupuesto																		
Análisis medioambiental																		
Conclusiones																		

Fig. 5.3.1. Diagrama de Gantt para el proyecto final de grado.

4. Introducción a los fundamentos teóricos

Para poder entender la metodología que se utiliza en el proyecto y los posteriores resultados, se debe hacer una breve, pero concisa introducción de los fundamentos matemáticos utilizados en los análisis realizados por el *software* R™. Como el trabajo aborda los métodos de análisis multivariante, se comenzará explicando los datos multivariantes, para proseguir con una introducción al análisis multivariante y los tipos que utilizaremos, el análisis de componentes principales (ACP) y el análisis factorial exploratorio (AFE).

4.1. Datos multivariantes

En el análisis multivariante (AM), la información de entrada es de carácter multidimensional ya que se busca interpretar los datos de más de una variable estadística sobre un conjunto de individuos [2]. Debido a que esta información multivariante con la que se trabaja es una matriz de datos *n x p*, debemos introducir los conceptos de algunos métodos algebraicos de especial relevancia en el AM.

4.1.1. Matrices de datos

Supongamos que sobre los individuos $\omega_1;\ldots;\omega_n$ se han observado las variables $X_1; \ldots ; X_p$. Sea $x_{ij} = X_j(\omega_i)$ la observación de la variable X_j sobre el individuo ω_i. La matriz de datos multivariantes es

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Las filas de **X** se identifican con los individuos y las columnas de X con las variables. Se indica:

1. $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_j, \ldots, \bar{x}_p)'$ el vector columna de las medias de las variables, siendo

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

2. La matriz simétrica *p x p* de covarianzas muestrales

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{pmatrix}$$

siendo

$$s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

Naturalmente, $\bar{x}$ y S son medidas descriptivas multivariantes de tendencia central y dispersión, respectivamente.

3. La matriz simétrica *p x p* de correlaciones muestrales

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

siendo $r_{jj'} = cor(X_j, X_{j'})$ el coeficiente de correlación (muestral). Este coeficiente viene dado por

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}$$

Donde $s_j, s_{j'}$ son las desviaciones típicas.

4.1.2. Variables compuestas

Los métodos de AM como el ACP y el AFE se basan en conseguir e interpretar combinaciones lineales adecuadas de las variables observables. Una variable compuesta *Y* es una combinación lineal de las variables observables con coeficientes $\mathbf{a} = (a_1, \dots, a_p)'$

$$Y = a_1X_1 + \cdots + a_pX_p$$

Si $\mathbf{X} = [X_1, \dots, X_p]$ es la matriz de datos, se puede escribir

$$Y = \boldsymbol{Xa}$$

Si $Z = b_1X_1 + \cdots + b_pX_p = \boldsymbol{Xb}$ es otra variable compuesta, se verifica:

1. $\bar{Y} = \bar{x}'\mathrm{a}$, $\bar{Z} = \bar{x}'\mathrm{b}$
2. $var(Y) = \mathrm{a}'\mathrm{Sa}$, $var(Z) = b'Sb$
3. $cov(Y, Z) = \mathrm{a}'\mathrm{Sb}$

Uno de los objetivos del análisis multivariante es encontrar variables compuestas adecuadas que expliquen aspectos relevantes de los datos.

4.1.3. Transformaciones lineales

Sea **T** una matriz *p x q*. Una transformación lineal de la matriz de datos es

$$\mathbf{Y} = \mathbf{XT}$$

Las columnas $Y_1, \dots, Y_q$ de **Y** son las variables transformadas.

4.1.4. Valores y vectores propios de una matriz

Es necesario explicar la obtención de valores y vectores propios debido a que en el análisis de componentes principales se utiliza para obtener dichos componentes.

Dada una matriz cuadrada A de orden *n* se dice que el número λ es un valor propio de A si existe un vector columna *c* no nulo tal que

$$Ac = \lambda c$$

El vector c se denomina vector propio de A asociado a su valor propio λ_0.

Para obtener estos valores y vectores propios se parte de la ecuación

$$Ac = \lambda Ic \quad ; \quad I = matriz\ identidad\ nxn$$

Es equivalente $(\lambda I - A)c = 0$ si *c* debe ser distinto de 0 necesariamente $|\lambda I - A| = 0$.

Se obtiene así el polinomio característico de A: $p(\lambda) := |\lambda I - A| = 0$. Resolviendo la ecuación de *n* grado, obtenemos diferentes valores propios λ. A continuación, resolvemos el sistema homogéneo indeterminado para un valor propio λ_n

$$(\lambda_n I - A)\begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

y obtenemos el vector propio correspondiente a ese valor propio λ_n.

4.2. Introducción al análisis multivariante

Como antes se ha comentado, en este proyecto se utilizarán dos tipos de análisis multivariante, ya que son los métodos que más encajan con los datos de interés. El ACP y el AFE son métodos de interdependencia, no distinguen entre variables dependientes e independientes y el objetivo de estas técnicas es identificar las variables que se relacionan, como se relacionan y porqué. En concreto, las dos técnicas tienen de entrada datos métricos.

Tanto el análisis de componentes principales como el análisis factorial se emplean para analizar interrelaciones entre un alto número de variables métricas definiendo estas

interrelaciones en términos de un número menor de variables que intentan explicar una elevada variabilidad reduciendo la dimensión. Las variables reducidas son denominadas componentes principales en el ACP y factores en el AFE.

4.3. Análisis de componentes principales

En el ACP se pretende estudiar las relaciones que presentan entre *p* variables correlacionadas. Para ello se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí, evitando la redundancia en la información. En esta transformación lineal (ver Capítulo 4.1.3.), las nuevas variables llamadas componentes principales, son combinaciones lineales de las antiguas variables que se van construyendo según el orden de las componentes que explicaran mayor variabilidad total de la muestra.

El único caso donde no tendría sentido realizar un ACP es si las variables iniciales ya están incorreladas de partida. En el caso de que los datos no cumplen normalidad multivariante se podría realizar la técnica matemática. No obstante, la normalidad multivariante de los datos puede dar una interpretación más precisa de dichos componentes.

4.3.1. Obtención de los componentes principales

Los componentes principales se obtienen a partir de la matriz de datos de entrada **X.** Estas componentes son las variables compuestas

$$Y_1 = Xt_1, \dots, Y_p = Xt_p$$

tales que:

1. $var(Y_1)$ es máxima a condición $t'_1 t_1 = 1$.
2. Entre todas las variables compuestas Y tales que $cov(Y_1, Y) = 0$, la variable Y_2 es tal que $var(Y_2)$ es máxima condicionado $t'_2 t_2 = 1$.
3. Si $p \geq 3$, la componente Y_3 es una variable incorrelacionada con Y_1,Y_2 con varianza máxima y análogamente se define la resta de componentes si $p > 3$.

Si $T = [t_1, \dots, t_p]$ es la matriz *p x p* cuyas columnas son los vectores que definen las componentes principales, entonces la transformación lineal $X \rightarrow Y$

$$\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} t_{i1} \\ \vdots \\ t_{ip} \end{pmatrix}$$

$$\boldsymbol{Y_i = XT_i} \;\; ; \;\; i = 1, \dots, p$$

se llama transformación por componentes principales.

Se indica:

1. Las variables compuestas $Y_i = Xt_i\,, i = 1, \dots, p,$ son las componentes principales.
2. Las varianzas son los valores propios de S

$$var(Y_i) = \lambda_i\,, i = 1, \dots, p.$$

3. Las componentes principales son variables incorrelacionadas:

$$cov(Y_i, Y_j) = 0\,, i \neq j = 1, \dots, p.$$

La primera componente Y_1 se obtiene de forma que su varianza sea máxima y sujeta a la condición de que la variable t_1 se encuentra normalizada. Para afrontar la problemática de maximizar con restricciones se aplica los multiplicadores de Lagrange, considerando la función lagrangiana:

$$L = t_1' S t_1 - \lambda(t_1' t_1 - 1)$$

Se deriva respecto a t_1 y se iguala a cero:

$$\frac{dL}{dt_1} = 2St_1 - 2\lambda t_1 = 0 \quad \rightarrow \quad (S - \lambda I)t_1 = 0 \qquad\qquad siendo\ I = matriz\ identidad$$

El sistema únicamente tiene solución si $|S - \lambda I| = 0$, que es equivalente a decir que λ es un valor propio de la matriz de covarianzas muestrales *S*. Por tanto, para maximizar $S(Y_1)$ hay que tomar el mayor valor propio λ de la matriz S, ya que $var(Y_1) = \lambda_1$. Tomando λ_1 como el mayor valor propio de S y t_1 como su vector propio normalizado asociado, se obtiene la primera componente principal, que se expresa:

$$Y_1 = Xt_1$$

Análogamente, la componente principal i-ésima se define como $Y_i = Xt_i$ donde t_i es el vector propio de S asociado a su i-ésimo mayor valor propio.

Si $m < p$ el porcentaje de variabilidad explicada por las *m* primeras componentes principales es

$$P_m = 100\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}$$

Cuando se aplica el método se desea que las primeras componentes principales describan un elevado porcentaje de la variabilidad total. Si esto sucede, se pueden sustituir las variables de entrada $X_1, \dots, X_p$ por las componentes principales $Y_1, \dots, Y_m$. En algunas aplicaciones,

estas componentes tienen interpretación empírica.

Existen diferentes métodos para determinar el número óptimo de componentes principales perdiendo la mínima información posible de los datos de entrada. En este trabajo se utilizará el criterio del porcentaje, ya que da más independencia al cliente de elegir el nivel de precisión que desea obtener. En el criterio del porcentaje se toma un número *m* de componentes principales de modo que P_m se aproxime a un valor especificado por el cliente. Un truco para observar número de componentes óptimos es observar si la representación de P_1, …, P_k, … con respecto de *k* se estabiliza a partir de una cierta componente principal *m*, por lo tanto, añadir dimensión no aportaría explicar más variabilidad.

Hay que remarcar, que el cálculo de los componentes principales si se hace con la matriz de covarianzas (como lo hace el *software* R™) depende de las unidades de medida empleadas. Si transformamos las unidades de medida, cambiarán a su vez los componentes obtenidos. [3]

Uno de los objetivos del cálculo de componentes principales es descubrir que información de la muestra describen. Sin embargo, normalmente es difícil interpretar estos resultados ya que se basan principalmente en explicar la mayor variabilidad posible y complican el poder etiquetar dichos componentes.

4.4. Análisis factorial

El AF persigue expresar *p* variables observables como una combinación lineal de *m* variables latentes. Estas variables de salida, denominadas factores, son hipotéticas y deben ser seleccionadas para explicar las intercorrelaciones entre las variables de entrada. En este método, las variables originales juegan el papel de variables dependientes que se explican por factores comunes y únicos, que son factores observables.

El AF puede ser exploratorio o confirmatorio. En el análisis factorial exploratorio no se conoce de partida el número de factores y es en la aplicación donde se determina. En el análisis factorial confirmatorio se suponen un número de factores de inicio y se intentan corroborar.

El análisis factorial obtiene e interpreta los factores comunes a partir de la matriz de correlaciones entre las variables ***R*** (ver Capítulo 4.1.1.)

4.4.1. El modelo multifactorial

El modelo de análisis factorial de *m* factores comunes considera que las *p* variables observables $X_1, \dots, X_p$ dependen de *m* variables latentes $F_1, \dots, F_m$, llamadas factores comunes, y *p* factores únicos $U_1, \dots, U_p$, de acuerdo con el modelo lineal:

$$X_1 = a_{11}F_1 + \cdots + a_{1m}F_m + d_1U_1$$
$$\ldots \qquad \ldots \qquad \ldots$$
$$X_p = a_{p1}F_1 + \cdots + a_{pm}F_m \qquad + d_pU_p$$

Las hipótesis del modelo son:

1. Incorrelación entre los factores comunes y los únicos dos a dos.
2. Los factores comunes están incorrelacionados con los factores únicos.
3. Los factores comunes y los únicos tienen media 0 y varianza 1 (variables reducidas).

En el modelo factorial, el conjunto de variables dependen de los factores comunes menos una parte de su variabilidad que es explicada por cada factor único correspondiente. La hipótesis 3 es una suposición teórica, ya que en general los datos observados no están reducidos.

El modelo factorial en expresión matricial es

$$\mathrm{X} = \mathrm{AF} + \mathrm{DU}$$

Indicamos por X el vector columna de las variables iniciales, F y U los vectores columna de los factores comunes y únicos respectivamente. D es la matriz diagonal con las saturaciones entre variables y factores únicos. En cuanto a la matriz factorial A, es una matriz *p x m* que contiene las saturaciones entre cada variable X y su factor F correspondiente.

$$\mathrm{A} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pm} \end{pmatrix}$$

El objetivo principal del análisis factorial es encontrar e interpretar la matriz A.

Podemos verificar de las condiciones del modelo

$$var(X_i) = (a_{i1}^2 + \cdots + a_{im}^2) + d_i^2 = (comunalidad) + unicidad$$

La comunalidad es la parte de la variabilidad de las variables solo explicada por los factores comunes y la unicidad la parte de la variabilidad explicada por el factor único correspondiente.

En el AF se puede estimar el número máximo de factores, para no sobredeterminar el modelo factorial se debe cumplir

$$m \leq \frac{1}{2}(2p + 1 - \sqrt{8p + 1}\,)$$

4.4.2. Método del factor principal

El planteamiento de los métodos para la extracción de factores se centra en que la matriz de covarianzas de las variables originales tipificadas $(X_1, \dots, X_p)$ es la matriz de correlación poblacional

$$R_p = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \qquad R_p = AA' + \Omega$$

Si se sustituye la matriz R_p por la matriz de correlación muestral R, los elementos de las matrices del segundo miembro de la expresión serán estimaciones en lugar de parámetros. Por lo tanto, $R = \hat{A}\hat{A}' + \hat{\Omega}$.

El método del factor principal es uno de los métodos para extraer los factores. En este método, se obtiene el primer factor maximizando la varianza explicada por él, es decir, maximizando $V_1 = var(X_1) = (a_{11}^2 + \cdots + a_{p1}^2)$ sujeta a las restricciones de la correlación muestral entre las variables X_h y X_j: $r_{hj} = \sum_{k=1}^{p} a_{hk}\, a_{jk}$. Para ello se pare del modelo factorial:

$$\begin{aligned} X_1 &= a_{11}F_1 + \cdots + a_{1m}F_m + e_1 \\ \dots \quad & \quad \dots \quad \quad \dots \\ X_p &= a_{p1}F_1 + \cdots + a_{pm}F_m + e_p \end{aligned}$$

Para resolver este problema de optimización con restricciones se utiliza el método de los multiplicadores de Lagrange, teniendo en cuenta la función:

$$G_1 = V_1 + \sum_{h,j=1}^{p} v_{hj}\,(r_{hj} - \sum_{k=1}^{m} a_{hk}a_{jk})$$

siendo v_{hj} los multiplicadores de Lagrange. Derivando la expresión para resolver el problema de optimización y realizando los ajustes necesarios en la expresión resultante, se concluye que λ_1 es el mayor valor propio de la matriz de correlaciones AA' y $(a_{11}, \dots, a_{p1})'$ es su vector propio asociado, de modulo λ_1. Por lo tanto, se tiene:

$a_{i1} = \alpha_{i1}\sqrt{\lambda_1}$ $\quad i = 1,2,3, \dots, p$ siendo $(\alpha_{11}, \dots, \alpha_{p1})$ un vector propio de modulo unidad.

Una vez obtenidos los pesos del primer factor se elimina su influencia considerando el siguiente modelo factorial:

$$X'_1 = X_1 - a_{11}F_1 = a_{12}F_2 + \cdots + a_{1m}F_m + e_1$$
... ...

$$X'_p = X_p - a_{p1}F_1 = a_{p2}F_2 + \cdots + a_{pm}F_m + e_p$$

Se encuentra el segundo factor maximizando la varianza explicada por él en el segundo modelo, que es $V_2 = var(X_2) = (a_{12}^2 + \cdots + a_{p2}^2)$ sujeta a las restricciones $r_{hj} = \sum_{k=1}^{p} a_{hk}\, a_{jk}$

Y como se concluyó con anterioridad:

$$a_{i2} = \alpha_{i2}\sqrt{\lambda_2} \qquad i = 1,2,3, \dots, p$$

λ_2 es el segundo mayor valor propio de la matriz de correlaciones AA' y $(\alpha_{11}, \dots, \alpha_{p1})$ un vector propio de modulo unidad.
Se continua con el proceso iterativo hasta obtener los pesos de todos los factores.
La matriz factorial será:

$$A = TD_{\lambda}^{1/2}$$

T = matriz cuyas k columnas son los vectores propios de AA' de módulo unidad.
D_λ = diagonal $(\lambda_1, \dots, \lambda_k)$

4.5. Diferencia entre el ACP y el AF

La diferencia general seria que en el análisis de componentes principales las variables de entrada explican las variables obtenidas, es decir, las componentes principales, y en cambio, en el análisis factorial los factores son los que explican las variables. Por ello, en el análisis factorial al suponer que existen unos factores que explican las variables son más fácilmente interpretables. El AF está orientado a analizar la covarianza y no la varianza total como en el ACP. En el análisis factorial se genera la hipótesis de que existen ciertas relaciones entre las variables y, por tanto, existen unos factores que pueden explicar gran parte de la comunalidad de los datos. En el ACP esta hipótesis no se genera y en el supuesto de que todas las variables iniciales estuviesen incorrelacionadas las componentes principales serían las mismas variables de entrada.

En resumen, las principales diferencias de la aplicación entre métodos es la mayor facilidad de interpretación de los factores que se obtiene del AF respecto al ACP y que puede hacer más sencillo el posterior análisis de resultados. Además, en el ACP escogiendo el suficiente número de componentes se explicaría la totalidad de su variabilidad y en el AF siempre queda una variabilidad no representada que es parte de la unicidad.

4.6. Análisis clúster

Para afinar aún más el análisis de los datos, se decide agrupar los individuos por similitud utilizando las variables de salida (factores o componentes principales) de los métodos multivariantes usados con anterioridad, así se podría clasificar los individuos según grupos que conforman una tendencia de compra. Este fin se realizará gracias al análisis clúster.

El análisis clúster, también conocido como análisis de conglomerados, es un método estadístico que busca agrupar los individuos tratando de lograr la máxima similitud entre elementos del mismo grupo y la mayor discrepancia entre los grupos. Los grupos homogéneos no son conocidos de antemano y no es necesario especificar un camino objetivo ajeno a la medida de las variables en los casos de la muestra de datos. Para poder discernir las diferencias entre grupos se debe emplear la distancia entre individuos. El cálculo de dichas distancias es un método objetivo a la hora de clasificar por grupos la muestra de datos. En el trabajo se usará el tipo de análisis de clúster no jerárquicos ya que los grupos que el propio análisis configura no dependen unos de otros. Y este tipo de análisis produce clusters disjuntos, en los cuales cada individuo pertenece solo a un grupo.

Una decisión importante antes de analizar por conglomerados es seleccionar las variables relevantes para identificar los grupos, como con anterioridad se habrá realizado un método de análisis multivariante se utilizarán esas variables de salida como las variables relevantes. Posteriormente, se debe seleccionar el criterio para hacer la clasificación de los individuos. En este trabajo se utilizará el algoritmo de *k-means*. Primero se debe saber el número de grupos que se quiere obtener y se parte de unas medias arbitrarias. Mediante pruebas sucesivas contrasta el efecto que sobre la varianza residual tiene la asignación de cada uno de los individuos a cada grupo.

El valor mínimo de varianza determina una configuración de nuevos grupos con sus respectivas medias. Se asignan otra vez todos los casos a estos nuevos centroides en un proceso que se repite hasta que ninguna transferencia puede ya disminuir la varianza residual. El procedimiento configura los grupos maximizando la distancia entre sus centros de gravedad. Para minimizar la varianza residual se debe maximizar la varianza intergrupal, ya que la varianza total es fija. Por lo tanto, la distancia euclídea al cuadrado es la utilizada por el método para conseguir variar la varianza intergrupal (denominada también factorial). La distancia euclídea se expresa con la siguiente formula:

$$d(x_i, x_j) = \sqrt{\sum_{c=i}^{p} (x_{ic} - x_{jc})^2}$$

Minimizar la varianza factorial es equivalente a conseguir que sea mínima la suma de distancias al cuadrado desde los individuos a la media del grupo al que van a ser asignados, es decir, minimizar la distancia euclídea al cuadrado.

Para cada media de arranque utilizada habrá una solución diferente, no obstante, el método lleva a unas clasificaciones muy similares o idénticas aun cambiando el valor aleatorio inicial.

Hay que resaltar que un problema importante para clasificar los datos en grupos es la elección de un número óptimo de grupos. La selección del número más apropiado al fenómeno que se analiza se basa en criterios tanto matemáticos como de interpretación. [4]

5. Resultados

A continuación, se expondrán los resultados de los diferentes tipos de análisis multivariantes sobre los datos de la facturación de una empresa distribuidora, estratificados por clientes y por las diferentes familias de productos que distribuyen. La empresa distribuidora es un mayorista de artículos de belleza.

5.1. Entrada de datos

La entrada de datos con la cual disponemos es una matriz de datos que sigue la siguiente estructura (figura 5.1.1.).

	Familia 1	…	…	Familia n
Código cliente 1	IMPORTES			
Código cliente m				

Fig. 5.1.1. Estructura de la matriz de datos de entrada.

Cada posición m x n tiene en cuenta la facturación de la familia n por el cliente m expresada en euros (€). El valor de facturación es el referido a un año natural, es decir, la facturación anual de cada familia por cliente. La dimensión de la entrada de datos es de 6115x15, se puede observar la necesidad de una reducción de la dimensión para poder hacer menos laborioso el trabajo de análisis de la facturación de los distintos clientes y poder actuar a partir del dicho análisis. En cuanto al código de cliente, cada comprador tiene un código exclusivo que lo identifica, es necesario diferenciar claramente cada comprador en la entrada de datos para posteriormente poder discutir la tendencia de compra de cada uno. Las facturaciones nulas de los clientes en las diferentes familias se expresan como una posición vacía y antes de realizar el análisis, si es necesario, se deberán convertir en 0.

Los diferentes tipos de familias identifican los tipos de productos, artículos o servicios que el proveedor puede proporcionar a sus clientes. Cada código numérico sirve para clasificar un conjunto de productos dentro de esa nomenclatura, a continuación, se relacionan esos códigos de familias con el correspondiente conjunto de artículos o servicios.

FAMILIA	NOMBRE
50	Peines y cepillos
100	Papel
150	Tijeras y navajas
200	Artículos eléctricos
250	Rulos, redes y gorros
300	Clips/Pinzas/Adornos
350	Capas/Tintura

FAMILIA	NOMBRE
400	Maquillaje/Manicura/Pedicura/Depilación
450	Mobiliario
500	Varios
510	Captain Cook (Marca popular)
550	Formación
600	Gran consumo
650	Cosméticos

Fig. 5.1.2. Código de familia de los diferentes tipos de productos.

Se hace una exploración estadística de los datos para tener una idea general del carácter y magnitud de los valores representados en la matriz. Con anterioridad, se han cambiado todos los espacios vacíos que son valores nulos de importe por el valor 0. Con esto, se consigue eliminar los posibles problemas de no tener representados los valores nulos en la misma unidad que los importes no nulos. En la figura 5.1.3. se observan los resultados de la exploración estadística.

	F050	*F100*	*F150*	*F200*	*F250*	*F300*	*F350*
Valor mínimo	0	0	0	0	0	0	0
Mediana	310	184	221	410	95	198	263
Media	1606	931	826	1390	518	888	926
Valor máximo	174988	557860	50173	298614	48269	42816	151054

	F400	*F450*	*F500*	*F510*	*F550*	*F600*	*F650*
Valor mínimo	0	0	0	0	0	0	0
Mediana	280	92	225	0	2	0	0
Media	992	1167	761	2	734	24	469
Valor máximo	252612	191801	97341	1774	66966	11377	171706

Fig. 5.1.3. Exploración estadística de los datos proporcionados (valores sin decimales).

Como se puede extraer de la exploración de los datos, la tendencia de los clientes en cada grupo es una compra de importe bajo o incluso nulo, exceptuando algunos casos que su valor es extraordinariamente elevado. Se observa que hay códigos de artículos que su mediana es casi nula, por lo tanto, sus valores estarán entorno al importe nulo. Para confirmar esta teoría, se realiza un histograma de los importes por código de artículo (figura 5.1.4.).

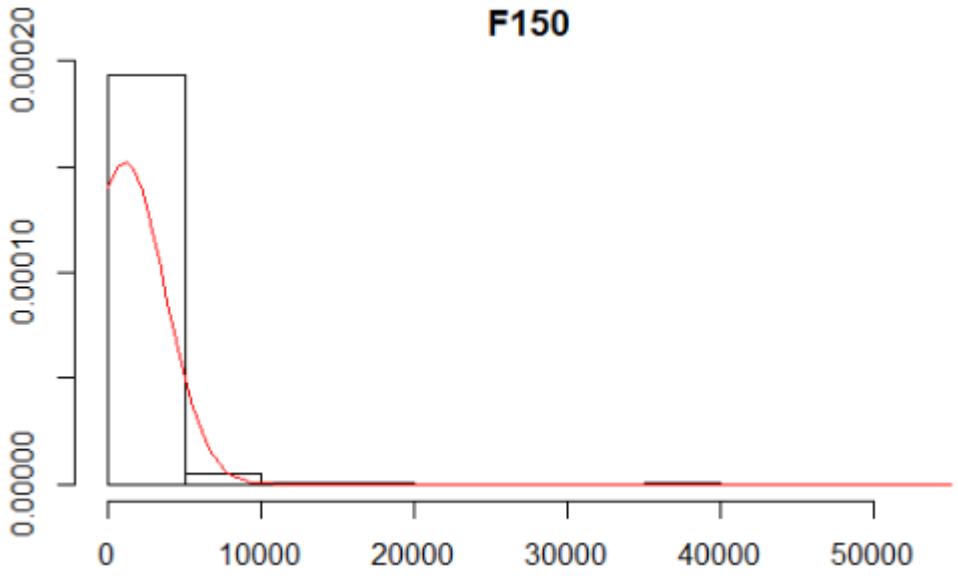

Fig. 5.1.4. Histograma de los importes vs densidad de datos estratificado por familia (En imagen F150).

En todas las familias se observa una distribución semejante a la vista en la figura 5.1.4. (se pueden encontrar todos los histogramas en el programa del anexo). La mayor densidad de datos está en los importes bajos rondando el 0 y el histograma sigue una tendencia

decreciente en todo momento, a excepción de algunos valores excepcionalmente altos. La normalidad multivariante no es una condición necesaria para hacer el análisis multivariante. No obstante, si se consigue normalizar los datos, se puede mejorar la interpretación de los factores.

Se decide hacer una transformación regida por el log(x+1) de cada importe para mejorar la normalidad de los datos. Se suma cada importe más uno para evitar los -∞ (menos infinito) que se ocasionarían por los importes nulos. Como se puede observar en la figura 5.1.5. se mejora la normalidad de los datos.

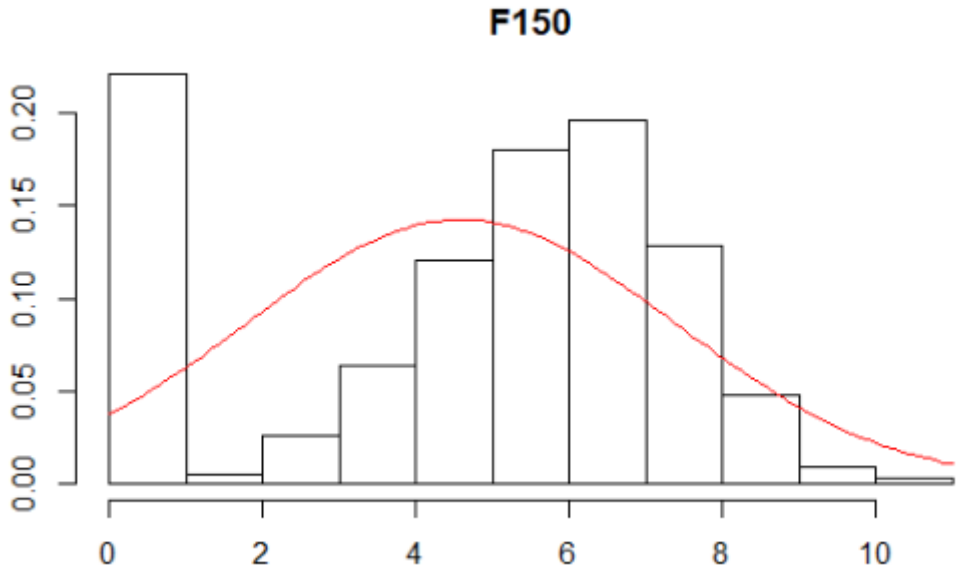

Fig. 5.1.4. Histograma del log(importes+1) vs densidad de datos estratificado por familia (En imagen F150).

Los valores nulos inflan el resultado de la transformación, aun así, se puede observar la normalidad de los datos. Con esta transformación ya se puede proceder a realizar los análisis multivariantes para intentar determinar la tendencia de compra de los clientes.

5.2. Análisis factorial

En el trabajo se realiza un análisis factorial confirmatorio, por lo tanto, se deben determinar los factores que se van a utilizar antes de realizar el análisis. Se decide utilizar un numero de factores menor que la mitad de las variables iniciales. Primero se prueba con 7 factores y de manera iterativa se prueba con 6. Se observa que la diferencia de varianza explicada es mínima, así que se cree idóneo usar 6 factores ya que es un valor que reduce lo suficiente la dimensión del problema y que puede explicar de forma correcta la tendencia de los clientes,

que es el principal objetivo del proyecto. En la figura 5.2.1. se observan los factores obtenidos del análisis factorial realizado con el programa R™, utilizando los fundamentos teóricos explicados en el capítulo 4.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
F050	0,716					-0,177
F100			0,722			
F150	0,342	0,145	0,485		-0,146	
F200		0,710				
F250	0,877					
F300	0,813					
F350	0,464		0,257			0,141
F400	0,275	0,334			0,350	-0,197
F450		0,384	0,135			0,362
F500	0,231		0,217		0,439	0,127
F510				0,669		
F550	0,446					0,283
F600				0,180	0,221	-0,145
F650	-0,137	0,511				-0,151

Fig. 5.2.1. Valores de los factores obtenidos (Los espacios vacíos son valores nulos o insignificantes).

En la figura 5.2.2. se recopila la variabilidad acumulada por cada factor, valor con el cual se puede estimar el grado de información explicada con la reducción de la dimensión.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Var. acu.	0,188	0,264	0,329	0,365	0,395	0,421

Fig. 5.2.2. Valores de la variabilidad acumulada por factor.

Como se puede observar, la variabilidad proporcionada por cada factor va disminuyendo progresivamente. A partir del sexto factor empieza a ser insignificante la información explicada, por ello puede que la variabilidad acumulada sea muy baja reduciendo la dimensión y por ello, la información explicada sea menor de lo esperado.

Posteriormente se representan los factores en un barplot para poder ver que factor representa cada tendencia de compra. Se analiza cada figura y así se contempla la tendencia de compra de familias de productos para cada factor.

- Factor 1: Este factor indica valores elevados en la facturación de las familias que incluyen productos como peines, cepillos, rulos, papeles, bolis, redes, gorros, pinzas y adornos (F050, F250 y F300). También se observa una ligera tendencia de compra de tijeras, navajas, capas, peinadores, tintura y el servicio de formación (F150, F350 y F550). El factor representa muy fielmente a los clientes que realizan compras referidas a los consumibles, es decir, a los productos básicos y perecederos a corto plazo. Se puede concluir que valores altos de este factor representa mayoritariamente al tipo de cliente que regenta una peluquería debido a la gran compra de productos perecederos y el uso del servicio de formaciones.

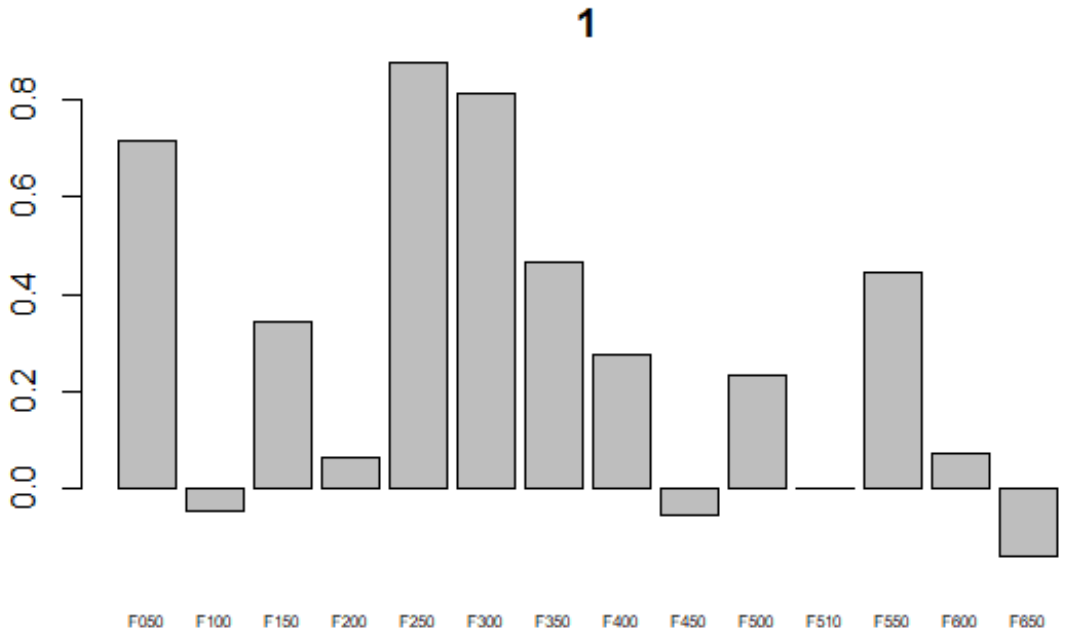

Fig. 5.2.3. Barplot del factor 1.

- Factor 2: Este factor indica valores elevados en la compra de artículos eléctricos y cosméticos (F200 y F650). También sugiere compras inferiores en maquillaje, manicura, pedicura, depilación y mobiliario (F400 y F450). Valores altos en este factor podría representar a los individuos que adquieren del proveedor los artículos de precio más elevado. La compra de estos individuos es temporalmente menos frecuente ya que el tiempo de vida de los artículos eléctricos es elevado y normalmente, los cosméticos son comprados en grandes lotes y son consumidos de una manera progresivamente lenta.

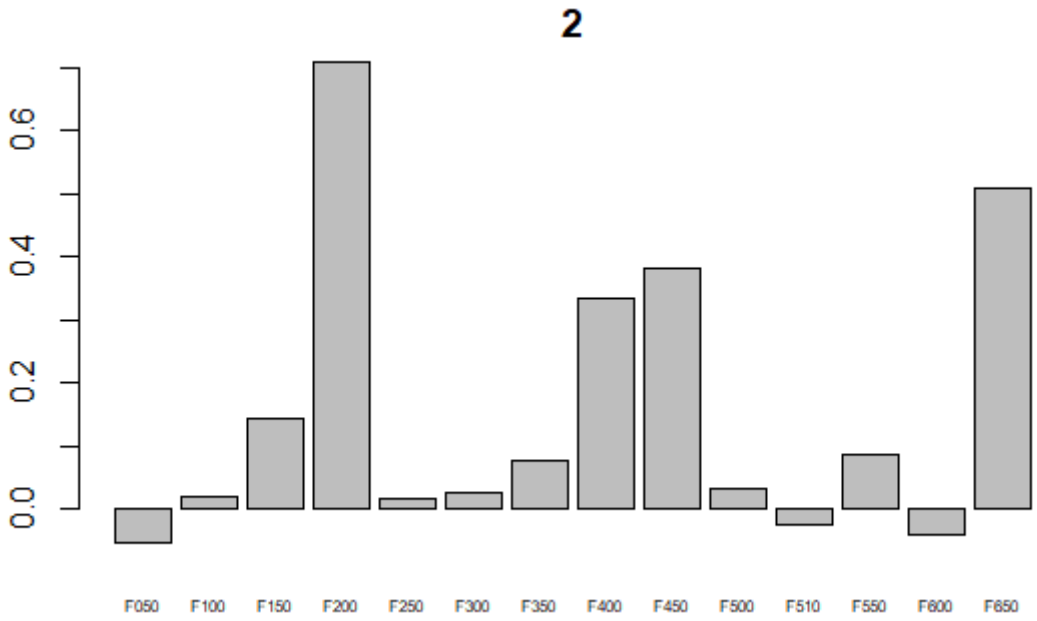

Fig. 5.2.4. Barplot del factor 2.

- Factor 3. Este factor indica valores elevados en la compra de papel, tijeras y navajas (F100 y F150). También se observa una ligera tendencia a la compra de capas, peinadores, tintura, mobiliario y varios (F350, F450 y F500). Valores altos en este factor representa cliente que compran papel en grandes cantidades.

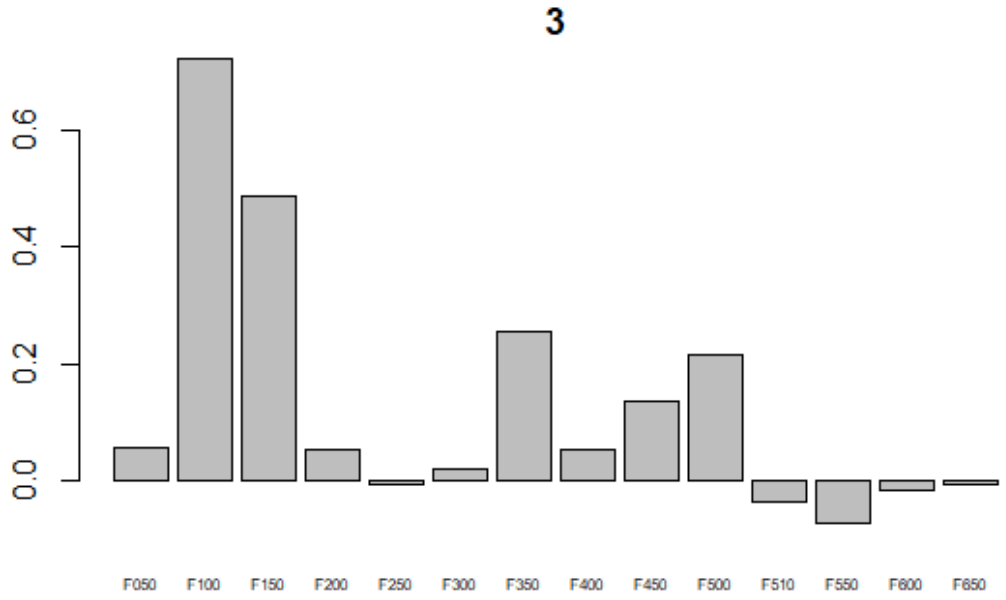

Fig. 5.2.5. Barplot del factor 3.

- Factor 4: Este factor indica valores elevados en la compra de productos de la marca Captain Cook (F510). Los productos de esta marca están destinados al público masculino, por lo tanto, valores altos en este factor caracteriza a barberías masculinas, peluquerías unisex o tiendas de cosméticos.

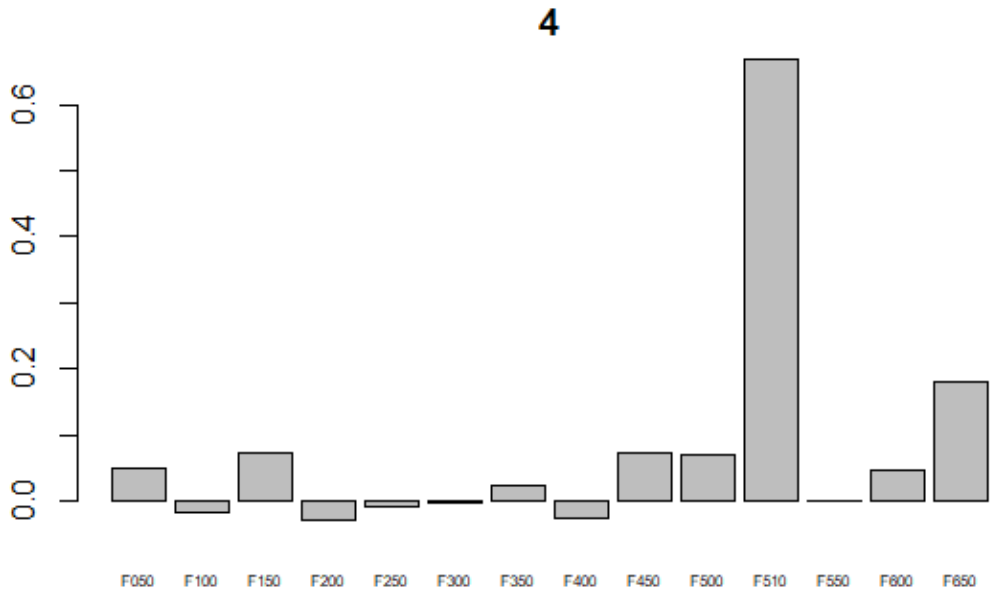

Fig. 5.2.6. Barplot del factor 4.

- Factor 5: Este factor indica valores elevados de compra de las familias de productos varios, maquillaje, manicura, pedicura, depilación y gran consumo (F500, F400 y F600). Valores altos de este factor representan a los clientes que compran productos de la gama de varios y de gran consumo. También se observa una tendencia negativa de compra de tijeras y navajas (F150), por lo tanto, se podría dar el caso de que los clientes fuesen salones de belleza que no incluyen el servicio de corte.

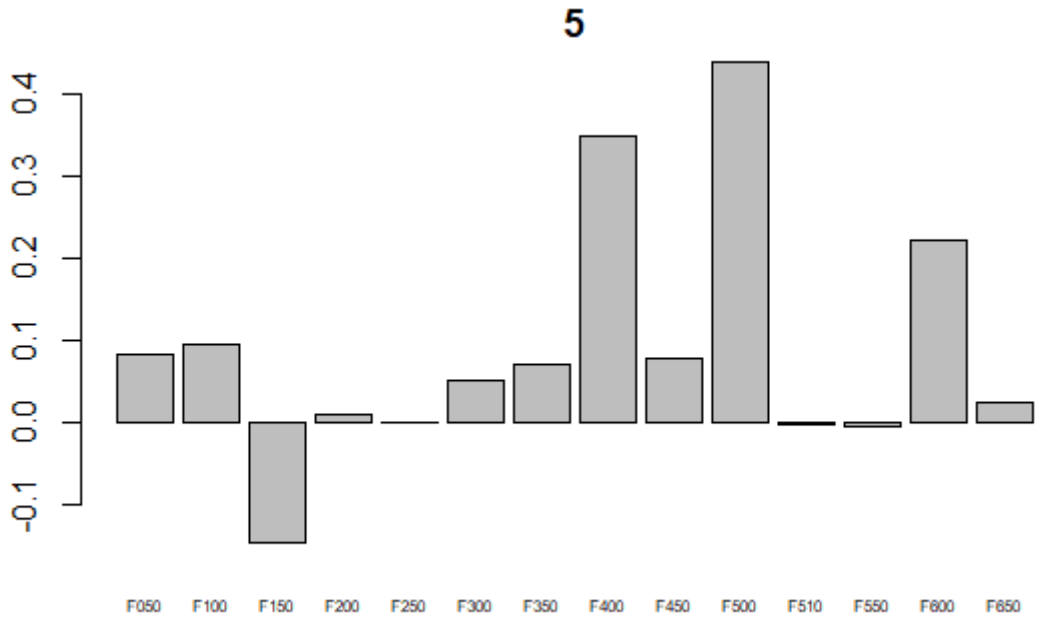

Fig. 5.2.7. Barplot del factor 5.

- Factor 6: Este factor indica valores ligeramente elevados de compra en las familias de mobiliario y formación (F450 y F550). Y se observa una tendencia negativa en la compra de peines, cepillos, maquillaje, manicura, pedicura, depilación, gran consumo y cosméticos (F050, F400, F600 y F650). Valores altos de este factor podrían caracterizar a los centros que ya tienen un alto stock de consumibles y utensilios, pero han tenido una remodelación tanto física como de la plantilla gastando parte del presupuesto en formación y mobiliario. También se puede deber al cliente con nuevas aperturas de peluquerías o salones de belleza.

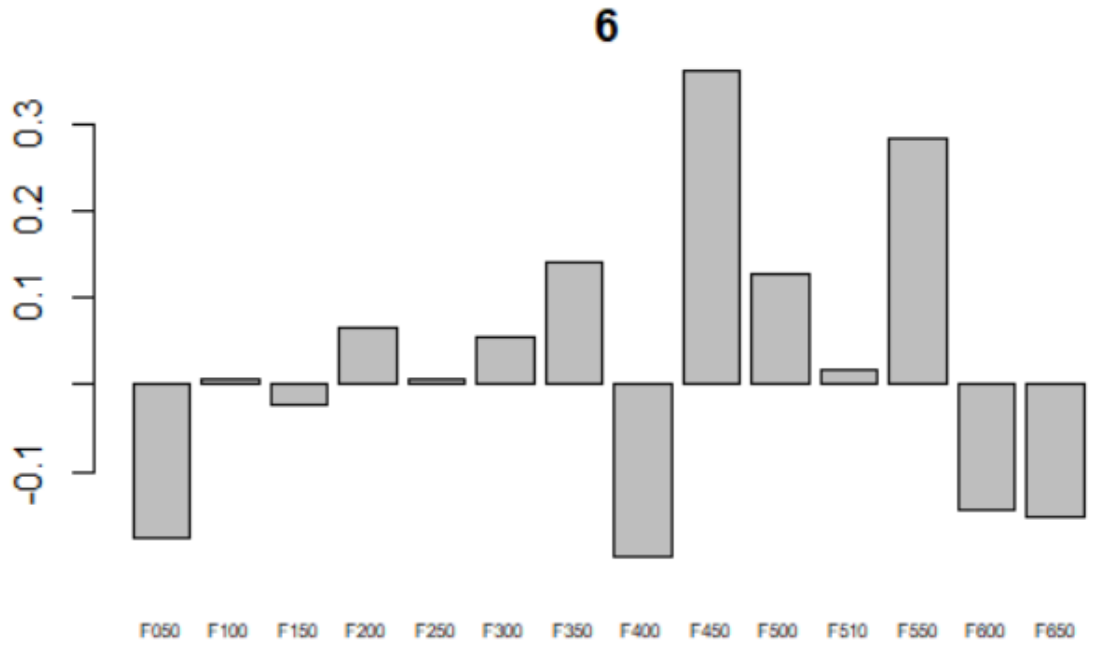

Fig. 5.2.8. Barplot del factor 6.

En el anterior análisis se puede determinar que factor caracteriza a las diferentes tendencias de compra que se pueden encontrar en el mercado de los productos de belleza.

Posteriormente, se realiza un análisis clúster con los factores obtenidos para así agrupar a los individuos en grupos afines y poder analizar la tendencia de comportamiento de los grupos en la compra. Como se vio en el capítulo 4.6, primero se debe escoger las variables de entrada que como se dijo con anterioridad serán los factores encontrados en el análisis factorial. El análisis por conglomerados (clusters) se realiza con el método k-means, y para determinar el número óptimo de centroides, equivalente a grupos, se analiza de manera iterativa cuando el total de la suma de cuadrados dentro de cada grupo, añadiendo centroides, para de decrecer de manera significativa. En la figura 5.2.9. se puede observar la gráfica del total de la suma de cuadrados dentro del grupo enfrente del número de centroides utilizado.

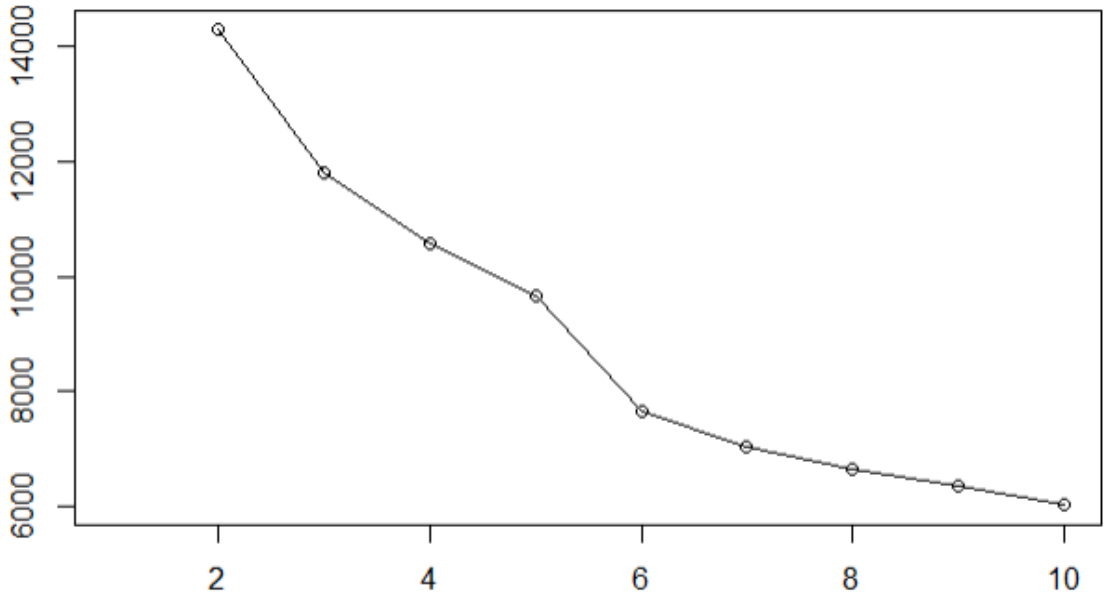

Fig. 5.2.9. Gráfica del total de la suma de cuadrados dentro del grupo vs número de centroides.

Como podemos ver a partir de 6 centroides (grupos) el descenso se reduce significativamente, por lo tanto, se podría afirmar que 6 grupos es un número correcto para el análisis de conglomerados que se realizará a continuación.

En la figura 5.2.10. se indica el número de individuos que conforman cada grupo calculado.

	G1	*G2*	*G3*	*G4*	*G5*	*G6*
Total de individuos	689	108	1277	1144	879	2017

Fig. 5.2.10. Total de individuos por grupo.

Los clientes se engloban bastante uniformemente entre los grupos, a excepción del grupo 2 que contiene muy pocos individuos y al grupo 6 que es el de más volumen de los seis.

Se grafican los individuos en diagrama bivariante donde se representan los diferentes valores de los factores que se enfrentan por cada individuo y estratificado por colores dependiendo del grupo que pertenezca cada cliente. En la figura 5.2.11. se incluye la leyenda de colores que se seguirá en los siguientes diagramas.

	Grupo 1	*Grupo 2*	*Grupo 3*	*Grupo 4*	*Grupo 5*	*Grupo 6*
Color	Rojo	Verde	Azul marino	Azul cian	Magenta	Amarillo

Fig. 5.2.11. Leyenda de colores para los diagramas bivariantes.

En las siguientes figuras se exponen los diagramas bivariantes.

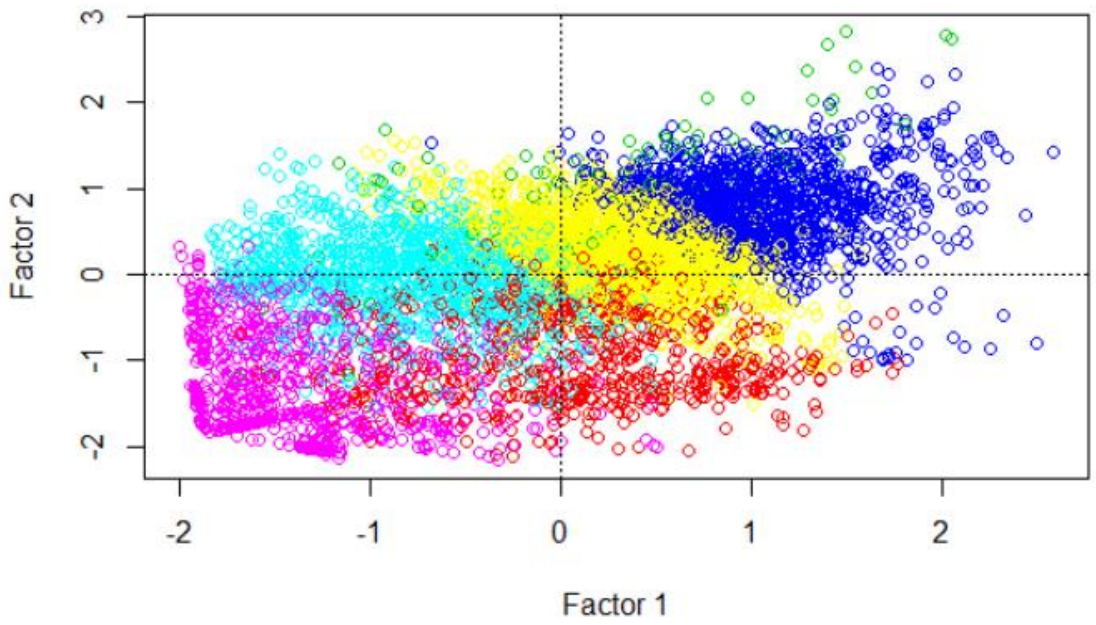

Fig. 5.2.12. Diagrama bivariante factor 1 vs factor 2.

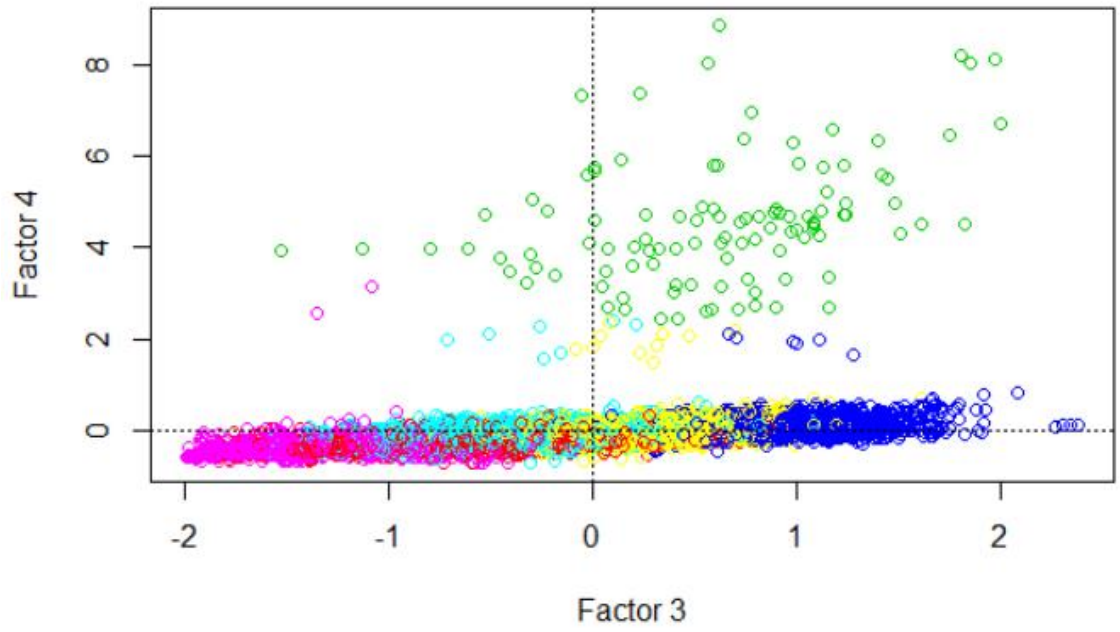

Fig. 5.2.13. Diagrama bivariante factor 3 vs factor 4.

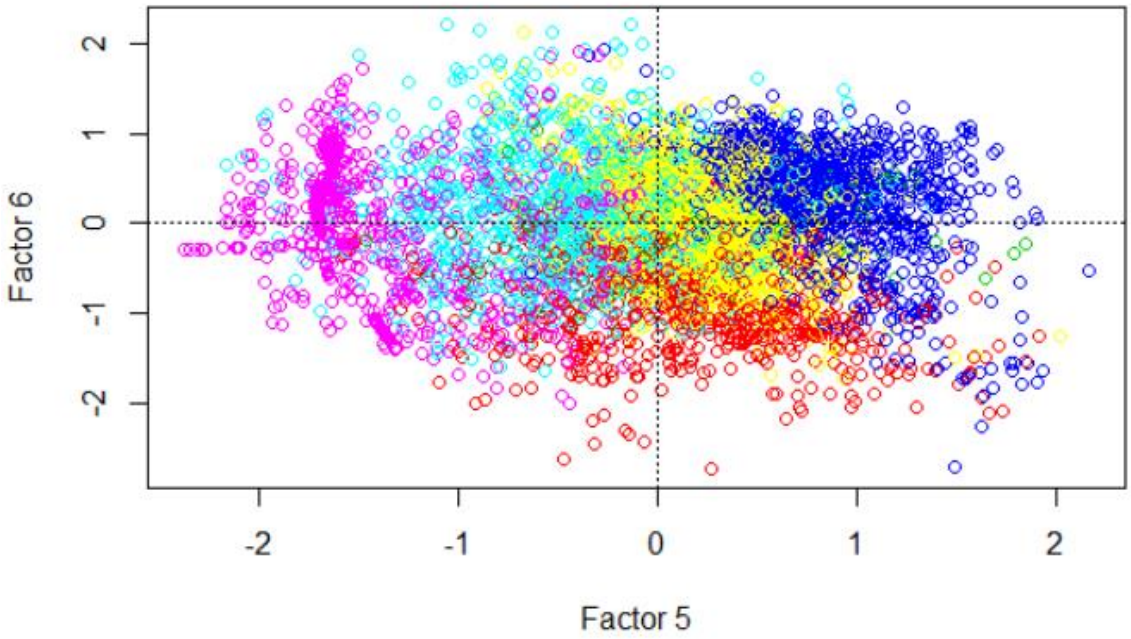

Fig. 5.2.14. Diagrama bivariante factor 5 vs factor 6.

A partir de los diagramas expuestos se identificará una posible tendencia de compra en cada grupo y dependiente de los resultados obtenidos en el análisis factorial.

- Grupo 1: Son clientes con valores negativos en el factor 2, es decir, consumen pocos artículos de vida útil elevada como podrían ser artículos eléctricos y cosméticos que se compran en grandes lotes. También se identifican con valores negativos del factor 3, compran poco papel, tijeras y navajas. Se observa valores ligeramente positivos en el factor 5, lo cual significa una compra de productos dedicados al sector de los salones de belleza como son artículos de maquillaje, manicura, pedicura y depilación. Se añade la gama de productos de gran consumo y los artículos varios, pero de compra moderada. En cuanto al factor 6, tiene valores negativos, por lo tanto, los importes de estos clientes en mobiliario y formación son muy bajos. La tendencia de los clientes englobados en el grupo 1 podría describirse como de empresas sin reestructuraciones próximas en el tiempo que basan su compra en artículos de belleza sin incluir herramientas para el corte de cabello.

- Grupo 2: A simple vista, el factor donde más predomina en valor positivo comparativamente respecto a los demás grupos es en el factor 4, esto indica que los clientes compran mucho de la marca Captain Cook, marca enfocada al público masculino. También se puede observar valores positivos del factor 3, por lo que indica una tendencia de compra alta de papel, tijeras y navajas. En los otros factores los resultados indican valores de compra significativos en la mayoría de gamas de productos exceptuando en formación y gran consumo. Del análisis de conglomerados se puede extraer que una gran mayoría de los individuos de este grupo pueden pertenecer al tipo de comercio de barbería masculina o peluquería unisex con alta clientela del genero masculino.

- Grupo 3: Se puede observar que en este grupo los valores de los factores 1,2,3,5 son positivos. Al analizar los resultados se puede afirmar que en este grupo los clientes compran mucho de los artículos perecederos, de los de gran vida útil como los aparatos eléctricos y de artículos de belleza básicos, como rulos, pinzas, adornos, etc. Se podría concluir con que los clientes pertenecientes a este grupo son empresas solidas en el sector que compran en gran cantidad.

- Grupo 4: Se puede observar que el grupo se centra en valores negativos de los factores 1 y 6, que están relacionados con valores de compra bajos en los artículos consumibles en un tiempo corto, y en mobiliario y formación. En este grupo se podrían englobar en clientes asentados en el sector que no compran al proveedor artículos perecederos ni tienen pensado ninguna remodelación

- Grupo 5: Los clientes de este grupo tiene valores negativos bajos de los factores 1,2,3 y 5. Estos valores representan una compra muy baja e incluso pudiendo ser nula de los artículos de la gama de consumibles, artículos eléctricos, cosméticos, maquillaje, varios y gran consumo. En este grupo se engloban los clientes que de la lista de facturación son los que menos compras compran en comparación con los demás individuos.

- Grupo 6: En este grupo los valores de todos los factores son cercanos al 0. Esto supone que los individuos de este grupo no tienen una tendencia marcada en la

compra de ningún producto. Son clientes de compra media que no destacan en ninguna gama de artículos ni servicios.

Analizando los resultados se obtiene que para cada factor del análisis factorial queda representado la compra alta o baja de las gamas de productos y con el análisis clúster, por cada grupo queda representado su respectivo valor de cada factor y así se puede razonar una tendencia de compra de cada grupo. Para poder ver la tendencia concreta de un cliente antes que observar todos los importes de cada familia existe la posibilidad de ver en qué grupo está y así tener una idea de su tendencia de compra.

5.3. Análisis de componentes principales

Se realiza el ACP mediante el programa R™ que utiliza los fundamentos teóricos explicados en el capítulo 4. Los componentes principales obtenidos por el programa están expuestos en la figura 5.3.1.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
F050	0,293	0,334	0,166		0,238	0,194	0,132
F100	0,287		0,135	-0,270	-0,753		-0,351
F150	0,330				-0,314	0,191	0,649
F200	0,278	-0,349				-0,789	0,293
F250	0,330	0,285			0,170		
F300	0,350	0,262			0,181		
F350	0,304						
F400	0,286		0,230		0,198	-0,242	-0,398
F450	0,269	-0,614	-0,355	-0,409	0,354	0,335	
F500	0,284						-0,410
F510							
F550	0,260		-0,769	0,515	-0,198		
F600		-0,461	0,405				-0,108
F650	0,136			0,686		0,329	

	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14
F050	0,341	0,210	0,601	0,343		0,145	
F100		0,320	0,145				
F150	0,351	-0,316	-0,291				
F200		0,134	0,237				
F250	-0,218	0,260	-0,110	-0,302	-0,695	-0,262	
F300	-0,347	0,195	-0,176	-0,380	0,627	0,205	
F350	-0,598	-0,354		0,634			
F400	0,410		-0,582	0,280		0,102	
F450		0,112					
F500	0,113	-0,695	0,294	-0,381			
F510							-0,999
F550	0,118						
F600	0,124				0,340	-0,919	
F650	-0,139						

Fig. 5.3.1. Valores de los componentes principales obtenidos (Los espacios vacíos son valores nulos o insignificantes).

Como se ha obtenido los catorce componentes principales tantos como variables existían en los datos de entrada se debe discutir con cuantos quedarse para explicar la mayor información posible reduciendo la dimensión del problema. Con los componentes obtenidos se intentará describir la tendencia de compra en los clientes. En la figura 5.3.2. se pueden observar la variabilidad acumulada añadiendo componentes principales.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Var. acu.	0,071	0,143	0,214	0,286	0,357	0,429	0,500

	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14
Var. acu.	0,571	0,643	0,714	0,786	0,857	0,929	1,000

Fig. 5.3.2. Valores de la variabilidad acumulada por componente principal.

Como se puede ver en la figura la variabilidad explicada proporcionada por cada componente es igual, por lo tanto, con la mitad de los valores únicamente se explicaría el 50% de la información contenida en la matriz de entrada. En el ACP se analiza toda la varianza tanto la común como la no común, eso explica el resultado [5]. En el capítulo teórico se explicó que para determinar el número de componentes optimo, el cliente y/o el que analiza los datos tiene que dar una variabilidad explicada de corte. Normalmente, se elige un valor entre el 70 y el 80% de variabilidad acumulada. Por todo ello, se decide escoger los 10 primeros componentes para el análisis y ver que tendencia representa cada componente principal mediante un barplot de cada componente. La reducción de la dimensión no será muy elevada ya que se pasa de 14 variables a 10. No obstante, la información que se pierde prescindiendo de esas variables será mínima.

- Componente principal 1: Esta componente se puede interpretar como medida de tamaño en el volumen de facturación. Valores altos en este componente representa una tendencia de compra del individuo centrada en todas las familias de productos exceptuando la marca Captain Cook y el gran consumo (F510 y F600). Es decir, el componente simula al cliente que compra de la mayoría de gamas de producto.

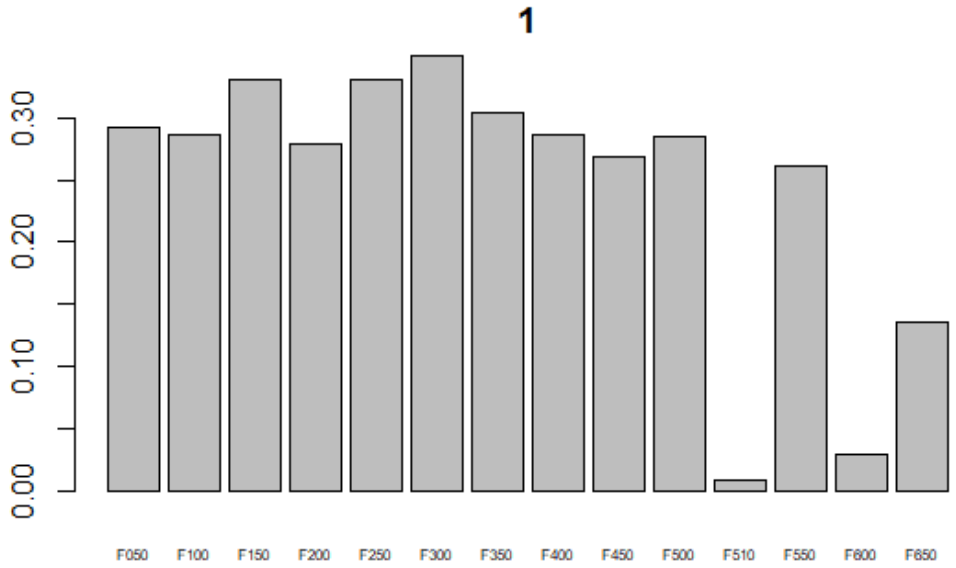

Fig. 5.3.3. Barplot del componente principal 1.

- Componente principal 2: Esta componente indica principalmente la tendencia negativa de compra en mobiliario, artículos eléctricos y cosméticos (F200, F450 y F650). También se observa una ligera tendencia positiva en la compra de peines, cepillos, rulos, redes, gorros, clip y pinzas (F050, F250 y F300). Valores elevados de esta componente podrían identificar a los clientes que compran habitualmente los artículos perecederos y no compran los de vida útil que probablemente ya dispongan.

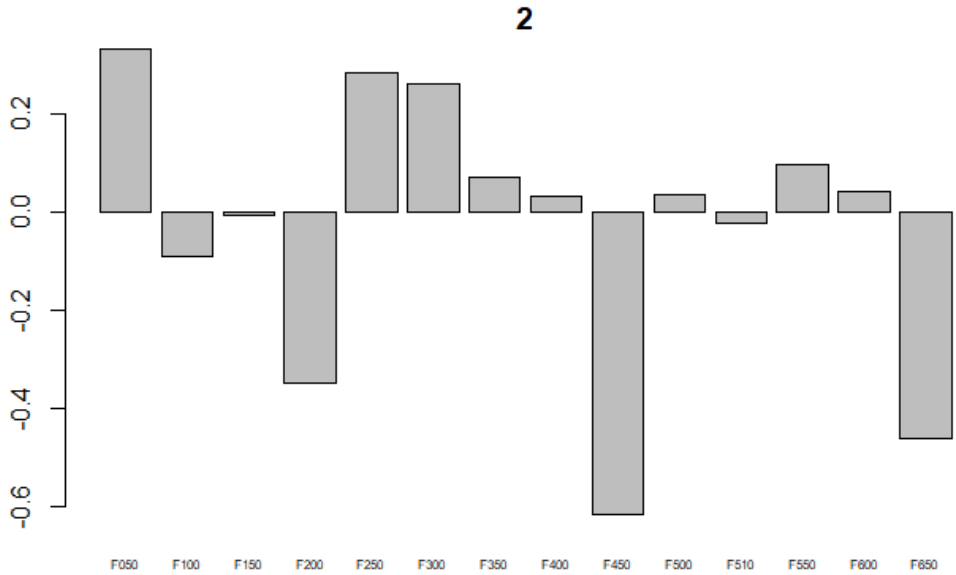

Fig. 5.3.4. Barplot del componente principal 2.

- Componente principal 3: Esta componente representa una gran tendencia negativa de compra en formación (F550) y una ligera tendencia negativa en mobiliario (F450). También se puede observar una ligera tendencia de compra de cosméticos. Valores altos de este componente podrían identificar clientes consolidados en el sector que no necesitan ni una ampliación de conocimiento en su plantilla ni mejora de su aspecto físico y funcional.

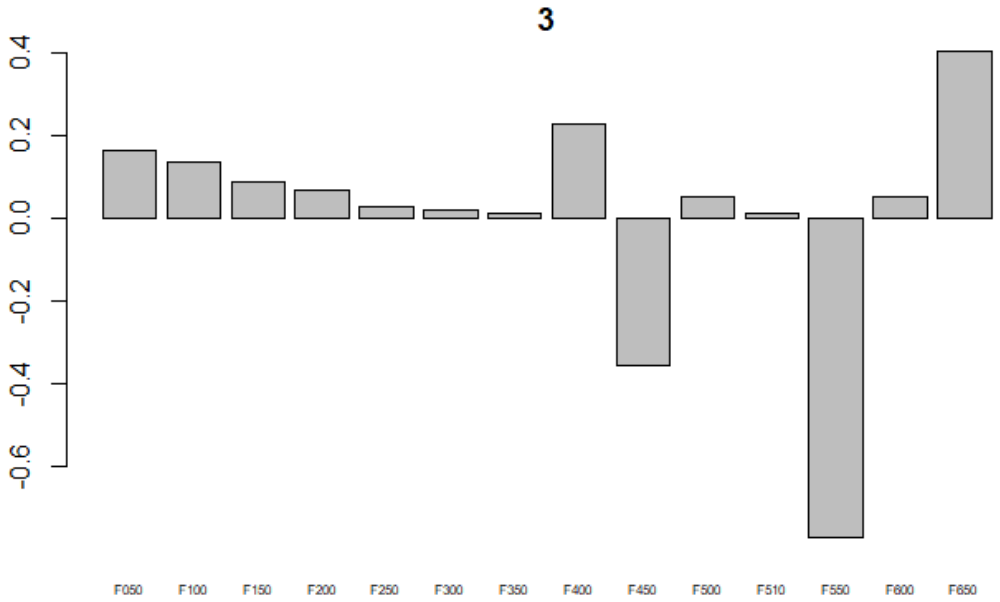

Fig. 5.3.5. Barplot del componente principal 3.

- Componente principal 4: Esta componente indica tendencia de compra positiva en cosméticos y formación (F550 y F650). También se observa tendencia negativa en papel y mobiliario (F100 y F450).

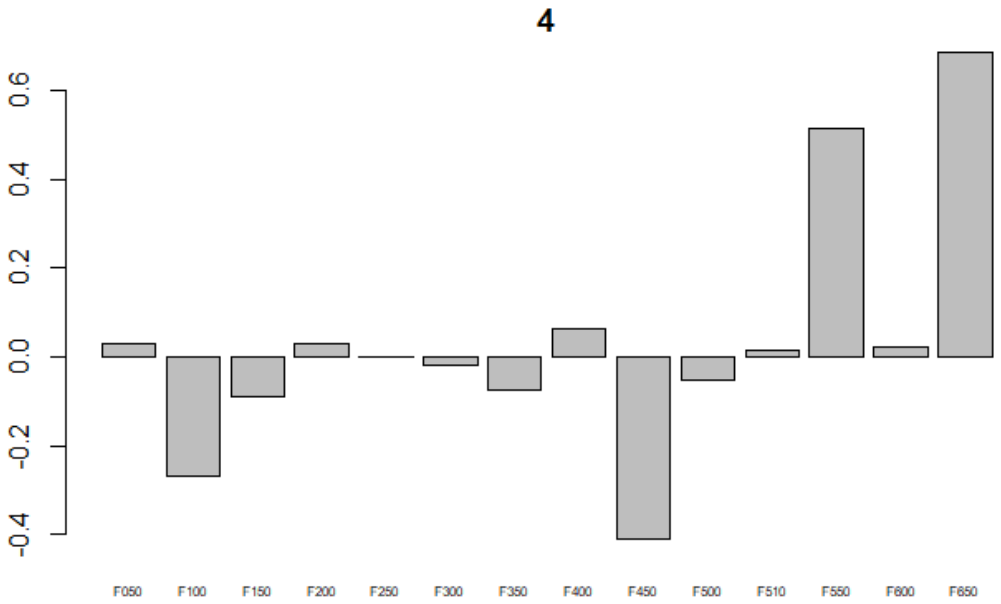

Fig. 5.3.6. Barplot del componente principal 4.

- Componente principal 5: Esta componente caracteriza la tendencia negativa de compra en papel (F100).

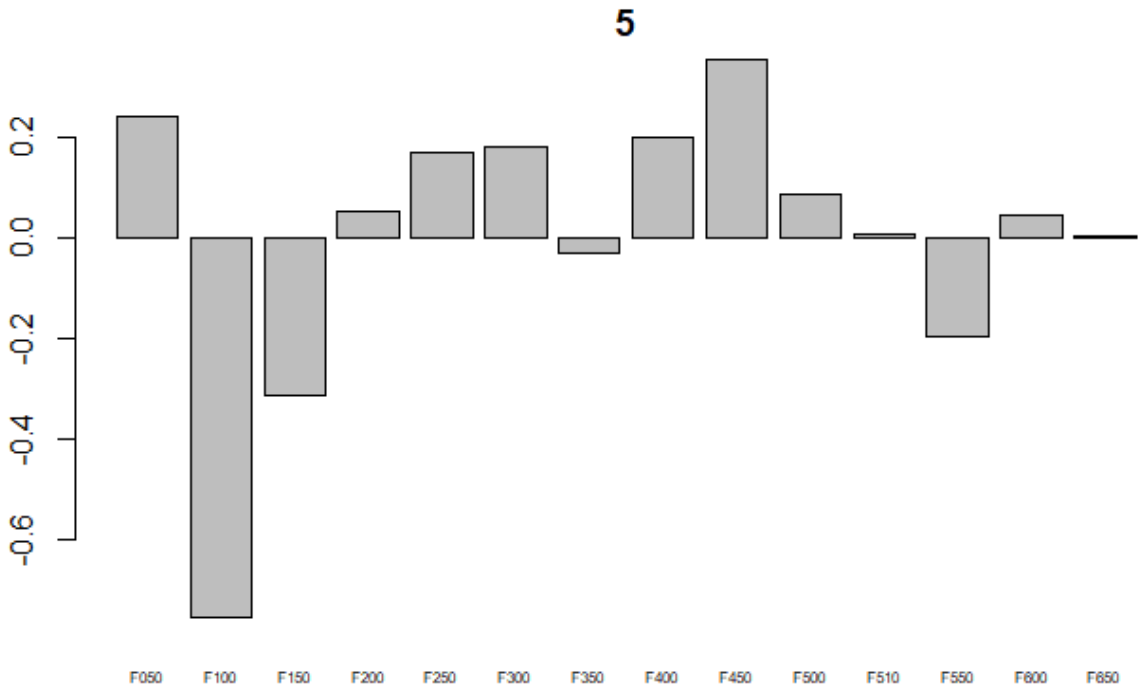

Fig. 5.3.7. Barplot del componente principal 5.

ETSEIB

- Componente principal 6: En esta componente se indica una gran tendencia negativa de compra en artículos eléctricos (F200). También se ve una ligera tendencia positiva de compra en mobiliario y cosméticos (F450 y F600), y una ligera tendencia negativa en maquillaje, manicura, pedicura y depilación (F400).

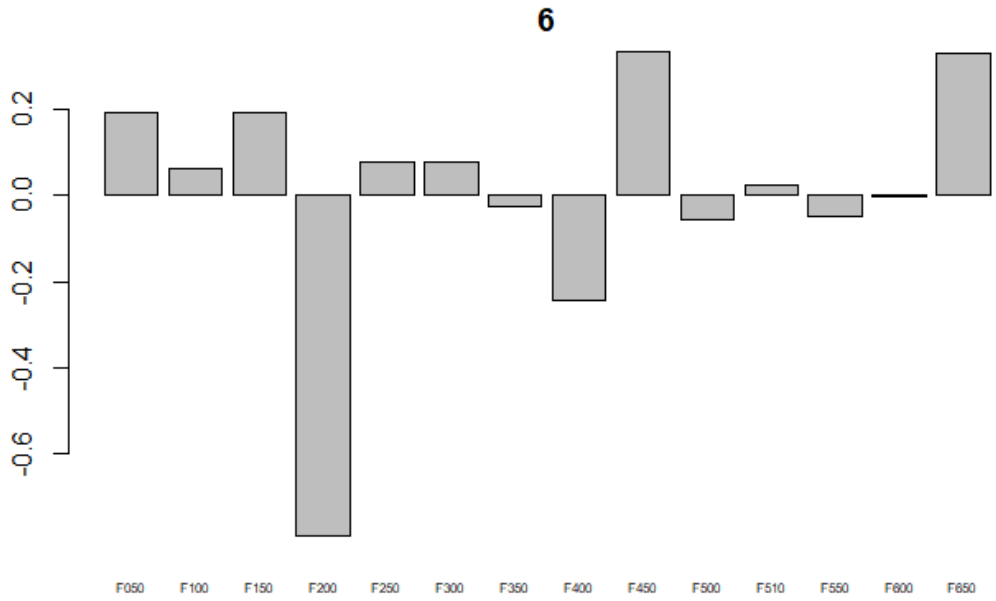

Fig. 5.3.8. Barplot del componente principal 6.

- Componente principal 7: En esta componente se indica la tendencia de compra de tijeras y navajas (F150). También se observa una ligera tendencia negativa de la compra de papel, maquillaje, manicura, pedicura, depilación y artículos varios (F100, F400 y F500).

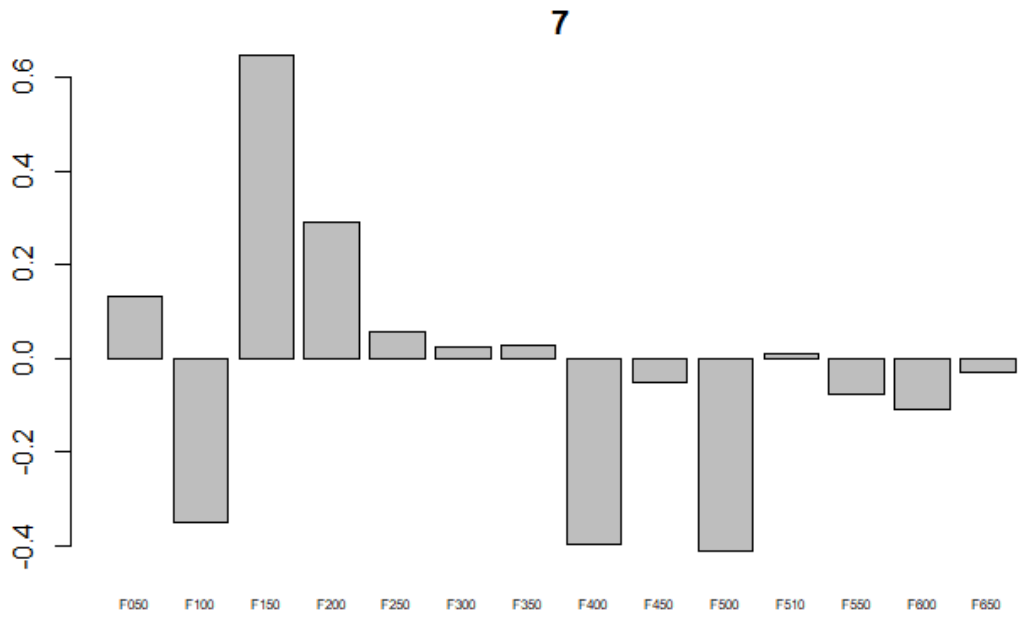

Fig. 5.3.9. Barplot del componente principal 7.

- Componente principal 8: En esta componente principal se destaca la tendencia negativa de compra de capas, peinadores y tintura (F350) y la ligera tendencia positiva en compra de peines, cepillos, tijeras, navajas, maquillaje, manicura, pedicura y depilación (F050, F150 y F400). También se observa una ligera tendencia negativa en la compra de rulos, redes, gorros, clips, pinzas y adornos (F250 y F300).

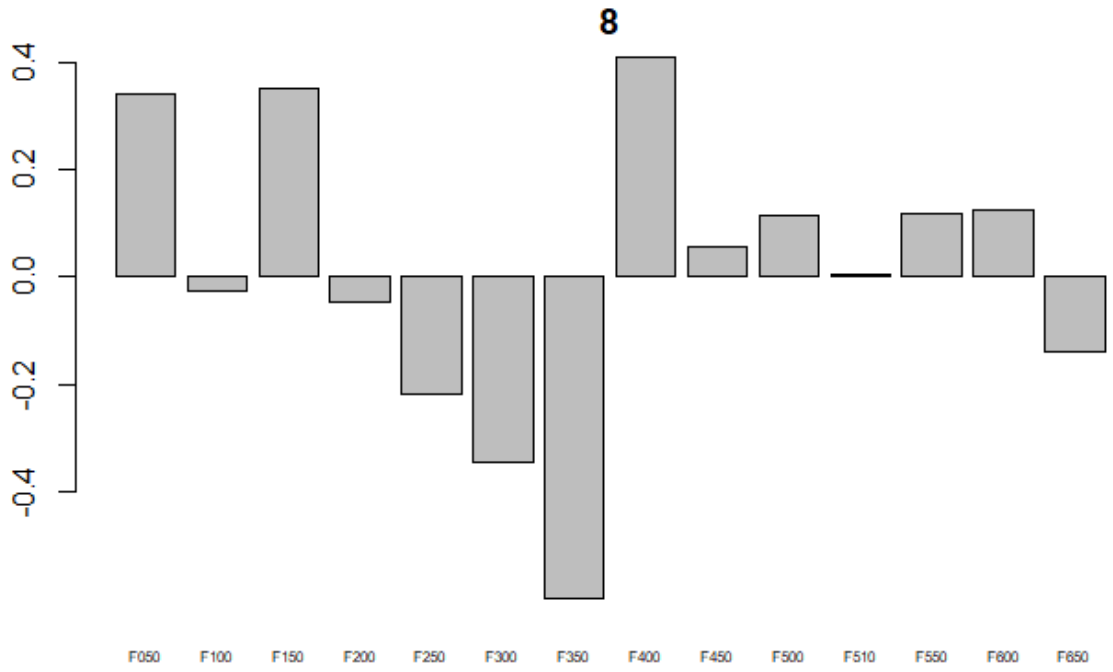

Fig. 5.3.10. Barplot del componente principal 8.

- Componente principal 9: En esta componente se destaca una gran tendencia negativa en la compra de artículos varios (F500). Se observa también una ligera tendencia negativa en la compra de tijeras, navajas, capas, peinadores y tintura (F150 y F350), y una ligera tendencia positiva en papel, rulos, redes y gorros (F100 y F250).

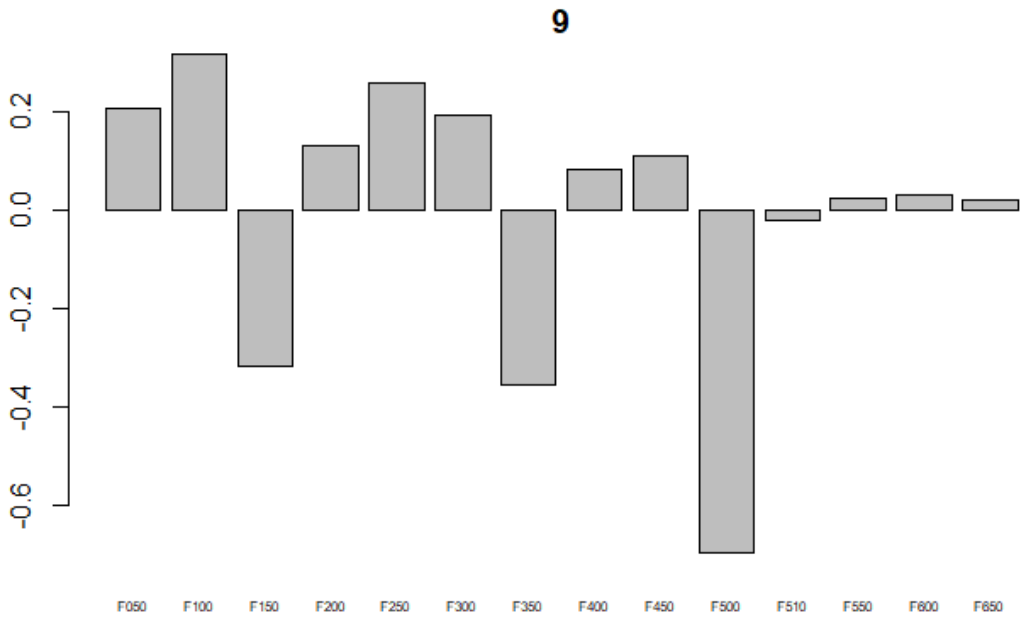

Fig. 5.3.11. Barplot del componente principal 9.

- Componente principal 10: En esta componente se destaca la gran tendencia positiva en compra peines y cepillos (F050) y la gran tendencia negativa en la compra de maquillaje, manicura, pedicura y depilación (F400).

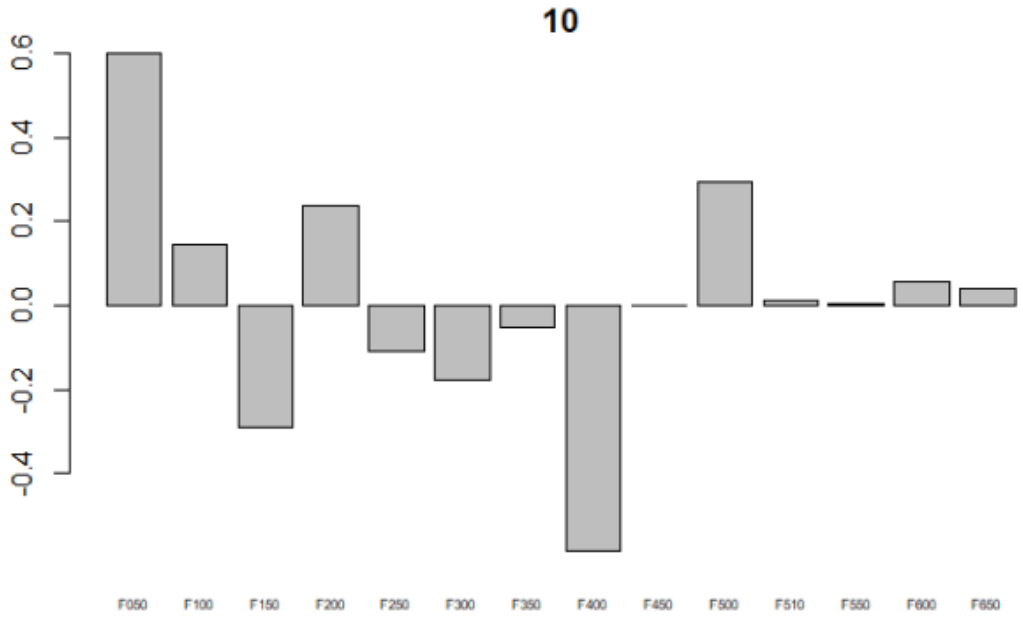

Fig. 5.3.12. Barplot del componente principal 10.

Se decide no realizar un análisis de conglomerados debido a que los componentes principales no son fácilmente interpretables y no iban a aportar mejores resultados que en el análisis factorial.
En este análisis de componentes principales se ha obtenido unos valores los cuales describen que gama de productos se compra más o menos. Cada cliente tendrá un valor de este componente y se podrá obtener una posible tendencia analizando la estimación de cada componente principal.

6. Presupuesto del proyecto

En este apartado se tendrá en cuenta el gasto total de la realización del proyecto, contando que quien realiza la tarea tiene la titulación de graduado en Ingeniería en tecnologías industriales. Primero de todo, se calcula el presupuesto inicial teórico para después compararlo con el real, teniendo en cuenta la variación de horas debidas a retrasos y problemas de diversa índole.

$$Presupuesto\ inicial = Horas\ invertidas\ x\ \left(\frac{salario}{h} + gasto\ eléctrico\right) + coste\ ofimática$$

$$Presupuesto\ inicial = 270\ x\ (14 + 0{,}072) + 40 = 3840\ €$$

A continuación, se desglosa el gasto obtenido con el tiempo real dedicado y los costes derivados de la realización del proyecto.

Presupuesto final total del estudio del proyecto			
Tarea / Concepto	*Horas dedicadas*	*Precio/Hora*	*Coste*
Salario proyectista	300 h	14 €/h	4200 €
Gasto eléctrico	300 h	0,6 kWh x 0,12 €/kWh	21,6 €
Licencia *Pack Office*	-	-	40 €
TOTAL	**300 h**	**-**	**4265 €**

Como se puede observar, el presupuesto real difiere alrededor de un +11% respecto al teórico. Este dato está dentro de los valores correctos ya que es permisible que el presupuesto real supere en un 10% al teórico.

7. Evaluación de impacto ambiental

La evaluación de impacto ambiental (EIA) es el procedimiento que se utiliza para identificar, evaluar y describir los impactos ambientales que puede producir un proyecto en el entorno en caso de ser ejecutado. En este proyecto no es un apartado de gran interés debido a que tanto la realización del proyecto como a su implementación si se diese el caso no afecta en medida significativa al entorno ni implica un peligro en la acción de preservar el medio ambiente. No obstante, se realizará la evaluación con el fin de tener más información para la posterior decisión de aplicar el proyecto en el mundo empresarial.

En este trabajo, el único factor que puede aportar un impacto negativo en el ámbito ambiental es el uso del ordenador y su correspondiente gasto eléctrico. En la siguiente formula se expresa el gasto total eléctrico consumido por las horas de realización del proyecto y de uso del ordenador.

$$Energia\ consumida\ PC = \frac{200W * 300h}{1000} = 60\ kWh$$

Si se busca en bibliografía [8], se puede encontrar datos sobre las emisiones de CO_2 por kWh de electricidad. El total de esas emisiones ocasionadas por la realización del trabajo se detallan a continuación.

	Energía final	Energía primaria	Emisiones
Electricidad	1 kWh	2,603 kWh	0,649 kg CO_2

$$Emisiones\ CO_2\ \text{totales} = 0{,}649\ \frac{\text{kg}\ CO_2}{\text{kWh}} * 60\ \text{kWh} = 38{,}94\ \text{kg}\ CO_2$$

Las emisiones totales son bastante bajas para el tiempo que dura el estudio y no debería considerarse una prioridad en cuanto a mejora ya que no supone un peligro para el medio ambiente. Además, este dato de emisiones se vería sustancialmente reducido una vez implementado el proyecto, ya que las horas de computación serían muy inferiores a las dedicadas en programación y análisis de los resultados.

Por lo tanto, se puede concluir que el impacto ambiental del proyecto y su posterior implementación es insignificante y no debería ser un factor prioritario a la hora de realizar el proyecto y ejecutarlo.

8. Conclusiones

En este apartado se dará respuesta a los objetivos SMART planteados al principio del proyecto, resumiendo los resultados obtenidos e interpretándolos cada uno en su contexto, llegando así a las soluciones más óptimas.

Resumen

El objetivo de este trabajo es analizar los dos métodos de análisis multivariante con los datos de entrada, el análisis factorial confirmatorio y el análisis de componentes principales, discutir las diferencias entre los dos métodos y elegir cual sería más conveniente para obtener la tendencia de compra de los clientes de esta empresa proveedora. Además, es interesante discutir si estos métodos de análisis pueden ser interesantes en otros casos para ver esa tendencia de compra y poder duplicar un análisis como el anteriormente hecho.

Análisis de los datos y decisión final

En primer lugar, se observan los resultados del análisis factorial. Para reducir de una manera significativa la dimensión de las variables se decide escoger 6 factores, pero con ello se pierde un 58% de la variabilidad explicada acumulada. Como se puede ver, los factores son muy interpretables y ya de por sí dan valiosa información del patrón de comportamiento que implicaría valores positivos o negativos de cada factor. Los factores obtenidos junto al análisis de conglomerados obtienen grupos donde se engloban los clientes por unas tendencias de compra muy definidas e interpretables.

En segundo lugar, en el análisis de componentes principales la adición componente proporciona la misma variabilidad explicada y para llegar a un nivel óptimo de información explicada reduciendo la dimensión se deben escoger 10 componentes principales. Se pasa de 14 a 10 variables una reducción de las variables muy pobre. En cuanto a la interpretación de los componentes principales, el único valor que es fácil de representar es el de volumen de compra, en los demás es compleja la visualización de un patrón de comportamiento.

Se puede analizar que las diferencias más significativas a la práctica entre el AFC y el ACP son:

- En el análisis factorial es mucho más fácil de interpretar y, por lo tanto, el análisis clúster obtiene grupos con una tendencia de compra muy marcada. Estos grupos serían de gran ayuda para la empresa proveedora para ver los patrones de comportamiento principales de su abanico de clientes.

- Con el análisis de componentes principales se puede llegar a explicar el 100% de la variabilidad a diferencia del análisis factorial que siempre queda un porcentaje de la variabilidad sin explicar. Como se ve en el proyecto, en el ACP se pueden conseguir cifras correctas de variabilidad explicada, eso sí utilizando un número alto de componentes principales.

En conclusión, **el método de análisis multivariante más conveniente para detectar la tendencia de compra de los clientes de la empresa proveedora de productos de belleza seria el análisis factorial confirmatorio**, ya que reducimos de manera considerable la dimensión del problema con unos factores fácilmente interpretables y formando unos grupos de clientes con un patrón de comportamiento en la compra bien definidos y extrapolables. No obstante, la información explicada por este análisis es baja debido a que los datos están inflados en importes nulos y la normalidad de los datos esta desplazada, reduciendo así la variabilidad explicada por cada factor.

En cuanto a si el análisis es replicable para otros datos de entrada, primero se debe encontrar y corregir lo mejor posible los datos de entrada para que aumentar la normalidad de los datos y que se consigan factores que expliquen un alto porcentaje de la información contenida en los datos de entrada. Si se consigue una buena corrección de los datos, el análisis multivariante es una gran herramienta para reducir la dimensión del problema sin poner en riesgo la información explicada de inicio, y, además, los factores son fácilmente interpretables y medibles respecto al comportamiento que pueden tener los clientes de diferentes empresas. Con estos resultados, se podrían iniciar tácticas de retención de clientes e incluso abrir departamentos de mejora en los productos y servicios más demandados y que tienen un peso en la facturación de la empresa elevado.

Mejoras en la metodología

Dado los resultados de variabilidad explicada por el análisis factorial se puede ver una limitación en cuanto al uso de los métodos de análisis empleados y los datos de entrada. Como se puede observar en la exploración estadística de los datos, un alto número de los importes por cliente son cero, por lo tanto, en la representación del histograma la normalidad esta desplazada. Una posible solución para corregir estos datos sería utilizar el análisis factorial inflado en ceros (ZIFA) que está avalado por numerosos estudios en cuanto a la mejora de la precisión y de la información explicada con datos de entrada inflado a ceros. [7]

Referencias

Referencias bibliográficas y web-gráficas

[1] RUBIO ARANDA, E. R.: *Aplicación de las técnicas de análisis multivariante a datos de concesionarios de automoción*. 2018. Proyecto final de carrera. Escuela técnica superior de ingenieros industriales (UPM).

[2] Salvador Figueras, M (2000): "Introducción al Análisis Multivariante", [en línea] 5campus.com, Estadística [http://www.5campus.com/leccion/anamul , 19 de enero 2019]

[3] MARÍN DIAZARAQUE, J.M.: Tema 3: Análisis de componentes principales, [en línea] UC3M, Estadística.

[4] DE LA FUENTE FERNANDEZ, SANTIAGO.: Análisis conglomerados. UAM 2011. Fac. Ciencias Económicas y Empresariales.

[5] DE LA FUENTE FERNANDEZ, SANTIAGO.: Análisis factorial. UAM 2011. Fac. Ciencias Económicas y Empresariales.

[6] Girones, Jordi.: "K-means algorithm", [en línea] uoc.edu, Estadística [http://data-mining.business-intelligence.uoc.edu/k-means , 23 de febrero 2019]

[7] Pierson, Emma & Yau, Christopher.: "ZIFA: Dimensionality reduction for zero inflated single-cell gene expression analysis" [en línea] Genome Biology [https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0805-z , 10 de marzo 2020]

[8] García, Xavier.: "Sabe usted por qué se impulsa la certificación energética", [en línea] certibcn.com, Impacto ambiental [http://certibcn.com/la-certificacion-energetica-y-las-emisiones-de-co2/, 23 de marzo 2020]

Bibliografía complementaria

[1] Knobl, Esteban.: "Objetivos SMART", [en línea] titular.com, [https://www.titular.com/blog/objetivos-smart-que-son-y-como-utilizarlos, 12 de enero 2019]

[2] PEREZ, CESAR.: Análisis multivariante de datos. Garceta 2004. Estadística.

[3] M. CUADRAS, CARLES.: Análisis multivariante. 2004. Estadística aplicada.